



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/706,198	11/03/2000	Marc Alexander Najork	9772-0298-999	6430

22879 7590 01/14/2005

HEWLETT PACKARD COMPANY
P O BOX 272400, 3404 E. HARMONY ROAD
INTELLECTUAL PROPERTY ADMINISTRATION
FORT COLLINS, CO 80527-2400

EXAMINER

GODDARD, BRIAN D

ART UNIT PAPER NUMBER

2161

DATE MAILED: 01/14/2005

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Applicati n No. 09/706,198	Applicant(s) NAJORK, MARC ALEXANDER	
	Examin r Brian Goddard	Art Unit 2161	

-- Th MAILING DATE of this communication appears on th cover sh t with the correspond nc address --
Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 15 July 2004.
2a) ☒ This action is **FINAL**. 2b) ☐ This action is non-final.
3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) 1-4 and 6-18 is/are pending in the application.
4a) Of the above claim(s) _____ is/are withdrawn from consideration.
5) ☐ Claim(s) _____ is/are allowed.
6) ☒ Claim(s) 1-4 and 6-18 is/are rejected.
7) ☐ Claim(s) _____ is/are objected to.
8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
10) ☒ The drawing(s) filed on 03 November 2000 is/are: a) ☒ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
a) ☐ All b) ☐ Some * c) ☐ None of:
1. ☐ Certified copies of the priority documents have been received.
2. ☐ Certified copies of the priority documents have been received in Application No. _____.
3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application (PTO-152) |
| 3) <input type="checkbox"/> Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)
Paper No(s)/Mail Date _____ | 6) <input type="checkbox"/> Other: _____ |

DETAILED ACTION

1. This communication is responsive to the Amendment filed 15 July 2004.
2. Claims 1-4 and 6-18 are pending in this application. Claims 1, 6 and 10 are independent claims. In the Amendment filed 15 July 2004, claim 5 was cancelled and claims 15-18 were added. This action is made Final.

Claim Rejections - 35 USC § 102

The text of those sections of Title 35, U.S. Code not included in this action can be found in a prior Office action.

3. Claims 1, 6 and 10 are rejected under 35 U.S.C. 102(a) as being anticipated by the article entitled "Mercator: A scalable, extensible Web crawler" by Heydon et al.

Referring to claim 1, Heydon discloses the method of downloading data sets by a plurality of web crawlers as claimed. See Figure 1 and Sections 3.1 – 3.8 for the details of this disclosure. Heydon teaches "a method [See Fig. 1] of downloading data sets by a plurality of web crawlers [worker threads] from among a plurality of host computers, comprising the steps of:

assigning a web crawler identifier [FIFO subqueue] to each one of the plurality of web crawlers [See Section 3.2, third paragraph];

for each respective web crawler:

downloading at least one data set [See Step 2] that includes addresses of one or more referred data sets;

identifying [See Steps 5-8] the addresses [URL(s)] of the one or more referred data sets, wherein each identified address includes a host computer identifier [host name (See Sections 3.2 & 3.8)];

for each identified address:

generating a representation [canonical host name / host name fingerprint] of the host computer identifier [See Sections 3.2 & 3.8];

determining a web crawler identifier [the particular worker thread's subqueue] to which the representation corresponds [See Section 3.2]; and

when the determined web crawler identifier is not assigned to the respective web crawler, sending the identified address [queuing the URL] to the web crawler to which the determined web crawler identifier is assigned [to the subqueue of the worker thread assigned to that host (See Section 3.2)]" as claimed.

Referring to claim 5, Heydon discloses the method for downloading data sets by a plurality of web crawlers as claimed. See Figure 1 and Sections 3.1 – 3.8 for the details of this disclosure. Heydon teaches "a method [See Fig. 1] of downloading data sets by a plurality of web crawlers [worker threads] from among a plurality of host computers, comprising the steps of:

for each respective web crawler:

receiving addresses [URL(s)] of one or more data sets from each of the plurality of web crawlers other than the respective web crawler [See Steps 8 & 1 and the discussion regarding claim 1 above];

for each received address:

determining [See Steps 4 & 7] if the address has been previously stored; and

if this determination is negative, storing the address [See Step 8]" as claimed.

Claim 6 is rejected on the same basis as claim 1. See the discussion regarding claim 1 above, as well as Figure 1 and the cited portions of the article for the details of this disclosure.

Claim 10 is rejected on the same basis as claim 1. See the discussions regarding claims 1 and 6 above for the details of this disclosure.

4. Claims 1-4 and 6-14 are rejected under 35 U.S.C. 102(e) as being anticipated by U.S. Patent No. 6,377,984 to Najork et al.

The applied reference has a common inventor with the instant application. Based upon the earlier effective U.S. filing date of the reference, it constitutes prior art under 35 U.S.C. 102(e). This rejection under 35 U.S.C. 102(e) might be overcome either by a showing under 37 CFR 1.132 that any invention disclosed but not claimed in the reference was derived from the inventor of this application and is thus not the invention "by another," or by an appropriate showing under 37 CFR 1.131.

Referring to claim 1, Najork discloses the method of downloading data sets by a plurality of web crawlers as claimed. See Figures 1-6 and the corresponding portions of Najork's specification for this disclosure. Najork teaches "a method [See Figs. 3 & 5-6]

of downloading data sets by a plurality of web crawlers [threads 130] from among a plurality of host computers, comprising the steps of:

assigning a web crawler identifier [queue identifier "i" (See Figs. 2-4)] to each one of the plurality of web crawlers [each thread (crawler) is assigned to exactly one queue (See Fig. 3B)];

for each respective web crawler:

downloading at least one data set [See Steps 334 & 560, and Column 4, line 63 et seq.] that includes addresses of one or more referred data sets;

identifying [See Steps 300, 500 & 564] the addresses [URL(s)] of the one or more referred data sets, wherein each identified address includes a host computer identifier [host name component "h"];

for each identified address:

generating a representation [canonical host name / host identifier "H"] of the host computer identifier [See Steps 301 & 502];

determining a web crawler identifier [See Steps 302-304, and 508 & 552] to which the representation corresponds; and

when the determined web crawler identifier is not assigned to the respective web crawler, sending the identified address [queuing the URL (See Steps 306, 510 & 554)] to the web crawler to which the determined web crawler identifier is assigned [See Figs. 3-5]" as claimed.

Referring to claim 2, Najork discloses the method of downloading data sets as claimed. See Figure 3 and the corresponding portion of Najork's specification for this

disclosure. Najork teaches the method of claim 1, as above, "wherein the plurality of web crawlers consists of n web crawlers [See Figs. 1-3]; and generating the representation includes computing a function [See Steps 302-304] of the host computer identifier [H] to generate an integer value [r] that is a member of a set of n predefined distinct values" as claimed.

Referring to claim 3, Najork discloses the method of downloading data sets as claimed. See Figure 3 and the corresponding portion of Najork's specification for this disclosure. Najork teaches the method of claim 1, as above, "wherein the plurality of web crawlers consists of n web crawlers [See Figs. 1-3]; and generating the representation includes computing a hash function [See Step 302] of the host computer identifier [H] to generate an intermediate value V [I], and computing V modulo n [See Step 304]" as claimed.

Referring to claim 4, Najork discloses the method of downloading data sets as claimed. See Figures 2-3 and the corresponding portions of Najork's specification for this disclosure. Najork teaches the method of claim 1, as above, "wherein the sending step includes: determining a web crawler address [r] for the web crawler [thread] to which the determined web crawler identifier is assigned [See Steps 302-306]; and transmitting the identified data set address [See Fig. 2] to the destination web crawler at the determined web crawler address" as claimed.

Referring to claim 5, Najork discloses a method of downloading data sets by a plurality of web crawlers as claimed. See Figures 1-6 and the corresponding portions of Najork's specification for this disclosure. Najork teaches "a method [See Figs. 3 & 5-6]

of downloading data sets by a plurality of web crawlers [threads 130] from among a plurality of host computers, comprising the steps of:

for each respective web crawler:

receiving addresses [URL(s)] of one or more data sets from each of the plurality of web crawlers other than the respective web crawler [See Fig. 2 and the discussion regarding claim 1 above];

for each received address:

determining [See Column 6, lines 48-52] if the address has been previously stored; and

if this determination is negative, storing the address [See the remainder of Fig. 5]" as claimed.

Claim 6 is rejected on the same basis as claim 1. See the discussion regarding claim 1 above for the details of this disclosure.

Claim 7 is rejected on the same basis as claim 3, in light of the basis for claim 6. See the discussions regarding claims 1, 3 and 6 above for the details of this disclosure.

Claim 8 is rejected on the same basis as claim 4, in light of the basis for claim 6. See the discussions regarding claims 1, 4 and 6 above for the details of this disclosure.

Referring to claim 9, Najork discloses the web crawler system as claimed. See Figure 4B and the corresponding portion of Najork's specification for this disclosure. Najork teaches the system of claim 6, as above, further comprising: for each respective web crawler, a lookup table [132]...as claimed.

Claim 10 is rejected on the same basis as claim 1. See the discussion regarding claim 1 above for the details of this disclosure.

Claims 11-13 are rejected on the same basis as claims 2-4 respectively, in light of the basis for claim 10. See the discussions regarding claims 1-4 and 10 above for the details of this disclosure.

Claim 14 is rejected on the same basis as claim 9, in light of the basis for claim 10. See the discussions regarding claims 9-10 above for the details of this disclosure.

5. Claims 1-4 and 6-14 are rejected under 35 U.S.C. 102(f) because the applicant did not invent the claimed subject matter. The claimed invention is fully disclosed in the article entitled "Mercator: A scalable, extensible Web crawler" by Heydon et al. and U.S. Patent No. 6,377,984 to Najork et al. as shown above. While applicant appears as party to both references (co-author of the article and co-inventor of the '984 Patent), at least one other author/inventor are party to each reference as well, showing that applicant did not invent the claimed subject matter alone.

6. Claims 1-4 and 6-14 are rejected under 35 U.S.C. 102(e) as being anticipated by U.S. Patent No. 6,182,085 to Eichstaedt et al.

Referring to claim 1, Eichstaedt discloses a method of downloading data sets by a plurality of web crawlers as claimed. See Figures 2-6 and the corresponding portions of Eichstaedt's specification for this disclosure. Eichstaedt teaches "a method of

downloading data sets by a plurality of web crawlers [gatherers (608)] from among a plurality of host computers, comprising the steps of:

 assigning a web crawler identifier [gatherer processor id "i" (See column 10)] to each one of the plurality of web crawlers;

 for each respective web crawler:

 downloading at least one data set that includes addresses of one or more referred data sets [See column 5, lines 35-50];

 identifying the addresses [URL(s)] of the one or more referred data sets, wherein each identified address includes a host computer identifier [host domain name (See Figs. 5-6 & columns 9-10)];

 for each identified address:

 generating a representation [superpage] of the host computer identifier;

 determining a web crawler identifier to which the representation corresponds [through mapping of superpages to gatherer processors (See Fig. 6)]; and

 when the determined web crawler identifier is not assigned to the respective web crawler, sending [forwarding/sending] the identified address to the web crawler to which the determined web crawler identifier is assigned [See column 6, lines 39-67 and column 12, lines 32-38]" as claimed.

Referring to claim 2, Eichstaedt discloses the method of downloading data sets as claimed. See Figure 6 and the corresponding portion of Eichstaedt's specification for this disclosure. Eichstaedt teaches the method of claim 1, as above, "wherein the

plurality of web crawlers consist of n $[k]$ web crawlers; and generating the representation includes computing a function [See Fig. 6 & corresponding portion of specification] of the host computer identifier [superpage] to generate an integer value [partition 606] that is a member of a set of n predefined distinct values [See Fig. 6]" as claimed.

Referring to claim 3, Eichstaedt discloses the method of downloading data sets as claimed. See Figure 6 and the corresponding portion of Eichstaedt's specification for this disclosure. Eichstaedt teaches the method of claim 1, as above, "wherein the plurality of web crawlers consists of n $[k]$ web crawlers; and generating the representation includes computing a hash function [communication hit-hash] of the host computer identifier [URL for an unknown superpage] to generate an intermediate value V , and computing V modulo n [See column 15]" as claimed.

Referring to claim 4, Eichstaedt discloses the method of downloading data sets as claimed. See column 6, line 47 – column 7, line 19 for the details of this disclosure. Eichstaedt teaches the method of claim 1, as above, "wherein the sending step includes: determining a web crawler address [Steps 404 & 406] for the web crawler to which the determined web crawler identifier is assigned; and transmitting [Step 405] the identified data set address [URL] to the destination web crawler [gatherer processor 403] at the determined web crawler address" as claimed.

Referring to claim 5, Eichstaedt discloses a method of downloading data sets by a plurality of web crawlers as claimed. See Figures 2-6 and the corresponding portions of Eichstaedt's specification for this disclosure. Eichstaedt teaches "a method of

downloading data sets by a plurality of web crawlers [gatherers (608)] from among a plurality of host computers, comprising the steps of:

for each respective web crawler:

receiving addresses of one or more data sets from each of the plurality of web crawlers other than the respective web crawler [See column 6, lines 39-67 and column 12, lines 32-38];

for each received address:

determining if the address has been previously stored [checks the already-visited pool (See Sections G – H in columns 12-15)]; and

if this determination is negative, storing the address [the URL is added to its local queue]" as claimed.

Claim 6 is rejected on the same basis as claim 1. See the discussion regarding claim 1 above for the details of this disclosure.

Claim 7 is rejected on the same basis as claim 3, in light of the basis for claim 6. See the discussions regarding claims 1, 3 and 6 above for the details of this disclosure.

Claim 8 is rejected on the same basis as claim 4, in light of the basis for claim 6. See the discussions regarding claims 1, 4 and 6 above for the details of this disclosure.

Referring to claim 9, Eichstaedt discloses the web crawler system as claimed. See Figure 4 and the corresponding portion of Eichstaedt's specification for this disclosure. Eichstaedt teaches the system of claim 6, as above, further comprising: for each respective web crawler, a lookup table [Tspace 406]... as claimed.

Claim 10 is rejected on the same basis as claim 1. See the discussion regarding claim 1 above for the details of this disclosure.

Claims 11-13 are rejected on the same basis as claims 2-4 respectively, in light of the basis for claim 10. See the discussions regarding claims 1-4 and 10 above for the details of this disclosure.

Claim 14 is rejected on the same basis as claim 9, in light of the basis for claim 10. See the discussions regarding claims 9-10 above for the details of this disclosure.

Claim Rejections - 35 USC § 103

The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

7. Claims 15-18 are rejected under 35 U.S.C. 103(a) as being unpatentable over Eichstaedt in view of Najork.

Referring to claim 15, Eichstaedt does not explicitly teach that each respective crawler includes multiple threads to download and process documents from a plurality of host computers as claimed.

Najork discloses a system and method similar to that of Eichstaedt, wherein a "web crawler" includes multiple threads to download and process documents from a plurality of host computers as claimed. See Figures 1-5 and the corresponding portions of Najork's specification for this disclosure.

It would have been obvious to one of ordinary skill in the art at the time the invention was made to modify each of Eichstaedt's crawlers/gatherers to include Najork's multiple threads to download and process documents from the plurality of host computers, to obtain the invention as claimed. One would have been motivated to do so to make each of Eichstaedt's gatherers more efficient, as disclosed by Najork.

Claims 16 and 17 are rejected on the same basis as claim 15, in light of the basis for claims 6 and 10 respectively. See the discussions regarding claims 1, 6, 10 and 15 above for the details of this disclosure.

Referring to claim 18, Eichstaedt v. Najork teaches the product of claim 17, as above, wherein each thread executes a main web crawler module [See Figs. 4-5] as claimed.

Response to Arguments

8. Applicant's arguments filed 15 July 2004 have been fully considered but they are not persuasive.

Referring to applicant's remarks on pages 7-9 regarding the Section 102 rejections of the independent claims over Heydon: Applicant argued that Heydon does not teach or suggest a plurality of web crawlers, and therefore does not teach "assigning a web crawler identifier to each one of the plurality of web crawlers," etc.

The examiner disagrees for the following reasons: Heydon's worker threads are each considered separate "web crawlers" as each thread performs the functions of a "web crawler." Nowhere has applicant provided a definition of "web crawler" that goes

above and beyond the conventional definition, that would distinguish from Heydon's worker thread. Further, applicant has not shown any specific difference between the claimed "web crawler" and Heydon's worker thread. Therefore, the Office interprets Heydon's plurality of worker threads as the claimed "plurality of web crawlers." Finally, Heydon does assign a web crawler identifier to each one of the plurality of web crawlers [worker threads] as the FIFO subqueue for each worker thread uniquely identifies that thread [crawler] within the system. Heydon teaches each and every limitation of the independent claims as shown in the prior Office action and repeated above.

Referring to applicant's remarks on pages 10-13 regarding the Section 102 rejections of the independent claims over Najork: Applicant argued that Najork does not teach or suggest a plurality of web crawlers, repeating substantially the same arguments as those directed to Heydon.

The examiner disagrees for the same reasons discussed above with regard to Heydon. Namely, each of Najork's worker threads is considered equivalent to a "web crawler" as claimed. Thus, Najork does teach a plurality of web crawlers, and each and every additional limitation of claims 1-4 & 6-14 as well.

Referring to applicant's remarks on pages 13-15 regarding the Section 102 rejections of the independent claims over Eichstaedt: Applicant argued that Eichstaedt does not teach or suggest "assigning a web crawler identifier to each one of the plurality of web crawlers" and further does not teach "generating a representation of the host computer identifier" and "determining a web crawler identifier to which the representation corresponds."

The examiner disagrees for the following reasons: Eichstaedt explicitly teaches assigning a web crawler identifier [i] to each one of the plurality of web crawlers [gatherers], generating a representation [superpage – part of sub-graph/partition] of a host computer identifier [host domain name], and determining a web crawler identifier to which the representation corresponds ['mapping' as shown in Fig. 6] as per the algorithm described for Figure 6. Applicant's arguments amount to nothing more than generic allegations that Eichstaedt is lacking certain claim elements. No logic or reasoning is provided to support these allegations. The portions of Eichstaedt's specification pointed out by applicant do not correspond to the portions cited in the grounds of rejection, showing a piecemeal analysis of the reference. Eichstaedt teaches each and every limitation of applicant's claims, and the rejection is therefore maintained and made Final.

Conclusion

9. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the

shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

10. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Brian Goddard whose telephone number is 571-272-4020. The examiner can normally be reached on M-F, 9 AM - 5 PM.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Safet Metjahic can be reached on 571-272-4023. The fax phone number for the organization where this application or proceeding is assigned is 703-872-9306.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

bdg
7 January 2005



SAFET METJAHIC
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 2100